The Art of Causal Conjecture

Review Author[s]:
Clark Glymour

Journal of the American Statistical Association, Vol. 93, No. 444 (Dec., 1998), 1513-1515.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199812%2993%3A444%3C1513%3ATAOCC%3E2.0.CO%3B2-R

Journal of the American Statistical Association is currently published by American Statistical Association.

The Art of Causal Conjecture.

Glenn SHAFER. Cambridge, MA: MIT Press, 1996. ISBN 0-262-1936-8X. xx + 511 pp. $50 (H).

Two centuries ago, except as an exercise in the application of combinatorics and in Jacob Bernoulli's inquiries into the foundations of inquiry, probability had almost no role in science. A century ago, probability had a very modest scientific role in statistical mechanics, in social statistics, and in Fechner's psychology. Today, scarcely a science is untouched by probability, either in the formulation of hypotheses or in their assessment in the light of data. It is proper that the sense of ideas so pervasive should now and then be reconsidered, and over this century there have been many reexaminations, from Borel, Ramsey, De Finnetti, von Mises, Kolmogorov, and others. Glenn Shafer's new book is a radical reconsideration of foundations.

Shafer's stalking horse is the sample space, random variable representation of probability that we owe chiefly to Kolmogorov. Kolmogorov's axiomatization has been influential not only because of its generality and mathematical precision, but also because of its neutrality. His axioms make no claim about the stuff of probability or how the theory is to be deployed, and so (almost) everyone can use them. Shafer does not reject Kolmogorov's formulation; he claims only that the axioms are not nearly sufficient. Many writers add something to Kolmogorov—frequentists, the idea of infinitely repeatable experiments; subjectivists, a further interpretation of conditional probability and a (somewhat equivocal) psychological reading of measures. Shafer adds the idea of a probability tree describing the possible processes through which the "events" represented in a sample space may come about. Whatever generality Kolmogorov's axioms gain by their spare and formal neutrality, Shafer establishes that a wealth of conceptual possibilities and interesting distinctions are lost by separating probability from the ideas about causal processes that sponsored its application in the sciences. Shafer's aim is nothing less than to unify probability and causation.

To contemporary readers, schooled more in measure theory than in the history and philosophy of science, Shafer's ambition may seem quixotic. It is not. The very idea of Bernoulli trials connects two ideas: absence of causal connection and probabilistic independence. After Bernoulli, who is Shafer's model, the first great scientific development of probability was the theory of the normal distribution, used as a justification of Legendre's method of reconciling discordant measurements by least squares, a method whose appeal lay both in its intuitive results and in its computational tractability. The derivation of the normal distribution had a gloss—the distribution is the approximate result of many independent small causes of either positive or negative deviation from a true value—that made the normal distribution a plausible treatment of errors of measurement in astronomy and geodesy. Later in the 19th century, the spread of probability to Galton's pseudobiology, to medicine, and to social statistics was typically, if not exclusively, in aid of resolving causal questions. The creation of psychometrics at the beginning of the 20th century used probability in causal theories of how the mind produces behavior, and later Fisher helped make probability a legitimate and essential part of biology. Fisher's influential development of the idea of randomized experiments extended the tie between probability and causality, although the discussion was oddly one-sided. His analyses brought mathematical methods to the assessment from experiments of hypotheses about probability, but left the essential connections with causation entirely informal.

Throughout the 19th and most of the 20th centuries, the mathematics and conceptual apparatus of probability and statistics became ever richer, while the causal ideas that drove much of that development remained tacit, informal, and obscure. Kolmogorov's axioms may be seen in retrospect as the completion of that tendency; all connection between probability and causal ideas is lost. No wonder, perhaps, that many statisticians since have treated ideas of causation as an embarrassing metaphysical entanglement, even while routinely practicing causal analysis without naming it. In the last 20 years, formal models of the relation between probability and causation have been developed, (the Rubin framework in statistics, and graphical causal models in statistics and in computer science), but they have met stiff resistance, sometimes more fervid than informed. The time seems right for Shafer's effort.

Shafer's book divides roughly into four parts. Chapters 1–4 introduce probability tree representations and develop an account of the meaning of probability. Chapters 5–13 use probability trees to define and relate a wealth of new ideas; for example, novel independence and conditional independence relations, and an equal abundance of novel association relations. Chapters 14 and 15 apply the apparatus to elucidate and distinguish the meanings of causation and to guide inquiry into causal relations, and chapter 16 considers a variety of graphical representations of causal processes. The final part of the book is a set of didactic appendices, useful in working through parts of the main text and often perceptive, but for the most part not intended to be original.

Shafer views Nature as the unfolding of a probability tree. The vertices of the tree are "situations." A situation, roughly, is a description of a possible state of the world at a time. (He does not say much about the root of the tree—presumably Creation, or the Big Bang, depending on one's theology.) At each vertex an "experiment" occurs, which may result, with various probabilities, in new situations that are the daughters in the tree of that vertex. Shafer calls the step from any vertex to one of its daughters [or sometimes a sequence of such steps (p. 45)] a "Humean event" (after the 18th century philosopher, David Hume). I confess to some confusion about the probabilities in Nature's tree. On one reading they are objective propensities—the probabilities will be 0 or 1 for deterministic transitions, or if there are genuinely indeterministic processes, as in quantum theory, something in between. On another reading, the probabilities in Nature's tree are epistemic; they measure the best possible predictions of an ideal observer, who is not omniscient. The steps in Nature's tree, the Humean events, are causes "of where we end up" (p. 9).

If we take a finite subtree of Nature's tree, then the terminal vertices (the leaves) of the finite tree that results may include many situations that have similar features; the roll of two die totals 7, for example. A set of events equivalent in some feature is a "Moivrean event." A suitable collection of sets of such vertices (a suitable collection of Moivrean events) is an ordinary sample space. What the tree represents, and the sample space does not represent, is the many different sequences of stages—(situations)—through which the Moivrean event (or its complement) might come about, and the changing probabilities of the Moivrean event in each possible stage. Each possible genesis of the Moivrean event is represented by a particular path from the root of the tree (or from any situation in which the Moivrean event has positive probability) to some vertex in the sample space event, and each path has a well-defined probability. (One die may collide with a side of the table and come up 6 while another does not collide with a side and comes up 1, or both may collide with sides and one come up 3 and the other 4, and so on. The reader can easily construct more interesting examples.)

Defining relations between tree features and sample space features allows one to define new relations among sample space events and random variables on the sample space. For example, a situation S is said to *resolve* a Moivrean (i.e., sample space) event E if the latter has either probability 1 or probability 0 at S, but does not have an extremal value in any ancestor of S (p. 37). E is determinate in any situation in which one of the ancestors resolves E. Of two Moivrean events, E, F, and E precedes F (or F is after E) if E is determinate in every situation in which F is determinate. Shafer has a lot more to say about Nature, but what about us?

Shafer says the probability tree "allows us to unify the subjective and objective aspects of probability in a single story about an observer" (p. 91). A limited, rational "observer," such as ourselves in our better moments, does not know Nature's probability tree, but in predicting and in explaining observations one may have an incomplete (or just inaccurate) version of Nature's tree with possibilities and probabilities all its own. The probabilities "describe the extent to which the observer is able to predict what will happen as events unfold, and they thereby tell us both about the beliefs of the observer and about her objective situation" (p. 91), and "at each step, the probabilities given by the tree are the best the observer can do in predicting what will happen next." The claim that the probabilities in an observer's tree "describe the extent to which the observer is able to predict" is immediately transformed into "probabilities define fair odds for bets" (p. 92). Shafer proves a connection between the long run expected values and probabilities as judged from any situation; it is almost certain that "subsequent events will happen in proportion to their average proba-

bilities" (p. 101), but he notes that this is a claim about the beliefs of an observer, not about what happens with what frequency.

What sense should we give to the claim that some real number between 0 and 1 expresses "the best the observer can do in predicting what will happen next"? What do "can" and "able" and "best" mean in Shafer's phrases? There is a straightforward sense in which an observer can beforehand assign 1 to the as-yet unknown actual outcome of a flip of a coin, and there is an elusive sense in which such an observer cannot. Shafer addresses these questions obliquely. He says the branching probabilities in a tree may be "confirmed" if the probabilities given by the tree match, on average, the frequencies the observer experiences (p. 106). This gloss has several problems. It does not ellucidate what proposition is being "confirmed" (confirmation so defined would seem to be a property of an entire tree, not of individual steps), and the observer only experiences (at most) one path through the tree. Shafer notes the latter point, which he calls Dawid's principle, but says it does not create a difficulty for the idea, which he calls "empirical relevance," that at each step the probabilities given by the tree are "the best the observer can do in predicting what will happen next" (p. 107). What follows is this:

> In order to test the empirical relevance of a particular probability tree, we must try out alternative methods of forecasting and see whether the probability trees they generate do any better.
>
> It is also important to acknowledge ... that validation is directed most fundamentally not to the probability tree but to the method by which it is constructed ... empirical validation of the method gives empirical meaning to paths not taken (p. 108).

The core of the matter, then, is that the sense in which the probabilities given by a tree are "the best the observer can do in predicting what will happen next" is that the tree and the probabilities are those that would be constructed from what the observer already believes by a method that is as good as any possible method similarly restricted. So what is a method; a function (from what?) to probability trees, or features of probability trees? Must the function be computable and, if so, how easily computable? How are methods to be compared for goodness? What are the trade-offs between goodness and computational tractability? How do we know that "best" makes sense, that there exists no better method? And why should the output of a method of inquiry and prediction have a probabilistic form at all, rather than, say, simply a prediction of the next situation, or the future sequence of situations, or a feature of that future sequence? These questions seem to me central to a genuinely foundational project, but aside from useful remarks on the variety of ways in which probability trees are implicitly or explicitly used, there is nothing more about them in the book. (There is a rich, relevant literature in computer science and elsewhere; some of the a-probabilistic literature has been reviewed and generalized in Kelly 1996, which I recommend to those who have given no thought to what inquiry would be like without probability.)

Shafer distinguishes several senses of independence. In an obvious notation, Moivrean events F and G are *formally independent* if for all situations S, $P_S(F \cap G) = P_s(F)P_s(G)$; *independent* if for each nonterminal situation S, either for all daughters T of $SP_T(F) = P_s(F)$ or for all daughters T of $SP_T(B) = P_s(B)$; and *weakly independent* if for each nonterminal situation S either $P(F)$ is the same in S as in the mother of S or $P(G)$ is the same in S as in the mother of S. Independence implies weak independence, which implies formal independence. None of the implications is reversible. The definitions generalize to any finite number of Moivrean events.

Notions of conditional independence, so important in usual causal analyses, also multiply. Moivrean events F and G are *formally independent posterior* to S if $P_T(F \cap G) = P_T(F)P_T(G)$ whenever T is "equal to or after S" (p. 128). F and G are *independent posterior* to S if they are independent in all situations equal to or after S (in an obvious specialization of the definition in the foregoing paragraph). F and G are *formally independent posterior to Moivrean event* E if $P_T(F \cap G) = P_T(F)P_T(G)$ in every situation T in which E has extremal probability. (Shafer says that E is *determinate.*) F and G are *independent modulo* E if there is no situation that influences both F and G without influencing E (where "influencing" means the probabilities change when moving from the situation to one of its daughters). F and G are *formally independent given* E if $P_T(F \cap G|E) = P_T(F|E)P_T(G|E)$ and, similarly for the complement of E, in every situation T that gives the complement of E positive probability.

One would expect a similar multiplication of kinds of association, and Shafer provides them. Moivrean event E *tracks* Moivrean event G in any two situations S and T such that $P_S(E) = P_T(E) = 1$ or $P_S(E) = P_T(E) = 0$, but not in the mother of S or of T (where S and T are situations where E *happens, or fails*), $P_S(G) = P_T(G)$. E is a *tracking positive sign* of G if E tracks G and $P_E(G) > P_{\sim E}(G)$, where $P_E(G)$ is the probability of G in situations where E happens and $P_{\sim E}(G)$ is the probability of G in situations where E fails. E is a *positive sign* of G if $P_S(G) > P_T(G)$ whenever $P_S(E) > P_T(E)$, and $P_S(G) < P_T(G)$ whenever $P_S(E) < P_T(E)$, where T is the mother of S. E is a *formal positive sign* of G if $P_S(G|E) > P_S(G| \sim E)$ in every situation S in which $P_S(E)$ and $P_S(\sim E)$ are both positive.

These various notions are of course related, and the book's middle chapters prove both general and special case connections. The definitions generalize naturally to relations among random variables defined on a sample space, including independence and uncorrelatedness, and an asymmetric notion that Shafer calls "unpredictability in mean"—a change in the probability distribution of one variable does not change the expected value of the second. In particular, a variable can track another variable and can be a sign of another variable.

Shafer connects these many distinctions (and more), and the metaphysical picture that they presuppose, with a variety of more familiar representations and discussions of causal relations. He includes an account of Martingales, hidden Markov models, still other graphical models, formal descriptions of probability trees, and a sensible (mostly—there are some falsisms among the truisms) analysis of maxims of inquiry. (One error should be noted: Shafer discusses Reichenbach's notion of a third event [or property] screening off the causal relation between two others, but misrepresents Reichenbach's idea; compare Reichenbach 1956, pp. 188–190 and Shafer pp. 162–165.) One of these topics is linear causal models of the kind common in social science, psychometrics, epidemiology, and elsewhere. Such models are sometimes represented graphically, although not generally as trees. The vertices of the graph are random variables, either with some substantive meaning or with "noises" or "errors." In Shafer's view, the underlying processes that such causal models purport to describe are features of a probability tree. Granting this, how much of the structure of such models, and of their sensible scientific and engineering use, can be accounted for in terms of probability trees? Shafer gives some interesting answers. The random variables in causal models are defined on a sample space of Moivrean events. In Shafer's view, there are in reality no causal relations among such variables, nor are there causal relations among the concrete events in which the variables take on values for units. The appearance of causal relations between, say smoking and cilia damage is epiphenomenal, the result of the unfolding of a probability tree whose steps are the actual causes of variables having whatever values they actually take on in a particular unit at a particular time. All genuine causal explanations are in terms of antecedent changes from one situation to another. When a linear causal model is appropriate, the specification in the model that $Z = bX + \varepsilon$, with the error term independent of X, may say that X is a linear sign of Z. All Humean events that result in an increase or decrease in the expected value of X are accompanied by an increase or decrease in the expected value of Z, always with the same constant of proportionality (p. 342); causally interpreted, the regression coefficient measures the difference in the expected value of Z between a situation in which no value of X happens, and a daughter situation in which some value of X happens (p. 313).

I do not know whether this is all that can be extracted from the probability tree framework in aid of explicating causal explanation, but I hope not. There are three related aspects of causal inquiry and causal explanation that Shafer's development of his framework does not yet engage, and in my view they are essential.

First, Shafer's reconstruction makes no use of the notion of interventions that alter the causal relations that would otherwise obtain. This is a signal virtue. But, even granting the correctness of Shafer's metaphysical picture, it is one thing to make no fundamental use of the notion of an intervention and quite another to make no connection with the notion, as Shafer does not. The decisions that we make in life, small or large, are themselves part of Nature's probability tree, but from the point of view of anyone making

or recommending an action, interventions are special; they are under the decision-maker's control.

Second, Shafer himself notes one of the limitations of his account of what causal models are about, accounting for what models are saying when they postulate mechanisms, or, slightly more formally, explaining the individual causal claims in causal models whose directed graph is multiply connected. Shafer points out that his treatment of linear signs, for example, is wholistic; causal relations among variables cannot generally be decomposed into parts in which some variables influence others both directly and indirectly through other variables: "the coefficients in a causal path diagram have a direct, but, in general, collective causal interpretation ..." (p. 344). The mechanisms postulated in the ordinary, shallow causal explanations of many sciences may very well be epiphenomenal or aggregates, but if correct, or even approximately correct, they nonetheless can provide either understanding or control or both.

Third, when for two variables there is (in ordinary parlance, not Shafer's) a third variable that influences both, the system is multiply connected. If the third variable is unobserved, then it is a latent variable, and also a confounder. Shafer's treatment of latent common causes of two observed variables is as measures of the dimensions of their actual common causes in the probability tree (p. 354). This is a nice idea, but it makes no connection with the importance of confounders in predicting the effects of interventions, new policies, or new experiments. Shafer notes that the analysis of causal relations as implying predictions about ideal interventions (Pearl 1995; Spirtes et al. 1993) does give an interpretation to complex mechanisms and to prediction of outcomes that, despite some vagaries, more or less accords with practice.

I do not believe that any of the aforementioned limitations are necessary consequences of the probability tree picture, and I certainly would not conclude that the fundamental notion of best prediction that can be made cannot be further clarified. Shafer's picture is sensible and suggestive; his development of it is original, brilliant, and fascinating. The worst one can say is true of even the greatest works: *The Art of Causal Conjecture* is incomplete, and what is missing matters.

Clark GLYMOUR
Carnegie Mellon University

REFERENCES

Kelly, K. (1996), *The Logic of Reliable Inquiry*, New York: Oxford University Press.

Pearl, J. (1995), "Causal Diagrams for Empirical Research," *Biometrika*, 82, 669–710.

Reichenbach, H. (1971), *The Direction of Time*, Berkeley, CA: University of California Press.

Spirtes, P., Glymour, C., and Scheines, R. (1993), *Causation, Prediction and Search*, New York: Springer-Verlag.

Exponential Families of Stochastic Processes.

U. KÜCHLER and M. SORENSEN. New York: Springer-Verlag, 1997. ISBN 0-387-94981-X. x + 322 pp. $54.95 (H).

Exponential families of distributions play a central role in statistical inference. Two types of exponential families are encountered in the statistical literature: noncurved and curved. Roughly speaking, the dimensions of the minimal sufficient statistic and the parameter space in the former are equal, whereas the dimension of the parameter space in the latter is smaller than that of the minimal sufficient statistic. The noncurved families possess various nice analytical and statistical properties and very often lead to explicit solutions in a variety of statistical questions. Each curved family can be embedded into a larger noncurved family; that is, it can be viewed as a subfamily of a noncurved family, derived via restrictions on the parameter space of the latter. The typical object in the statistical literature on exponential families in the classical iid setting is either a noncurved family or a curved family whose parameter space lies in the interior of the natural parameter space of a larger noncurved family. As a general rule, the noncurved families are nice and tractable, the curved families with the aforementioned property are not nice but still somewhat tractable, and the curved ones whose parameter space lies entirely on the boundary of the natural space of a larger noncurved family are much less tractable.

It turns out that the typical object in the theory of exponential families of stochastic processes is related to the latter case. A more specific explanation of this relationship follows. Note first that an exponential family of stochastic processes is a collection (indexed by the time parameter) of exponential families with a common parameter space. It rarely happens that the latter families can be embedded into noncurved families whose common (for different time epochs) parameter space interior contains the parameter space of the original exponential family of stochastic processes. But if this does happen for a family of stochastic processes, then their likelihood functions can be viewed as arising from models generated by processes with independent increments. The classical theory of exponential families is directly applicable to such models; Wiener, Poisson, and other important Lévy processes fall into this category. However, if such processes are observed in random time intervals, then the corresponding sequential likelihood functions, with the exception of those based on a few nice stopping rules, do not have this property. In other words, the sequential likelihood functions of these processes are typical objects, in the context mentioned earlier, in the theory of exponential families of stochastic processes. Therefore, it should come as no surprise that sequential methods play an important role in this theory.

Another important area is semimartingale theory, which has an established place in statistical inference for stochastic processes. Properties of the processes can imply restrictions on the type of possible exponential families associated with them, and, conversely, a specific type of exponential family may imply important probabilistic properties of the associated processes. The latter is an important feature that in particular identifies the theory of exponential families of stochastic processes as a possible tool for solving problems in other areas.

Exponential Families of Stochastic Processes is the first book treatment of this area, and it covers the progress made during the last decade. Both authors are leading experts in the field. The presentation is mathematically rigorous, and the exposition is clear and concise. I found only a few typographical errors.

The first three chapters are introductory. They contain formal definitions, examples, and more detailed treatment of Lévy processes, whose one-dimensional distributions belong to natural exponential families. Chapter 4 is devoted primarily to the property that a special exponential structure implies independence of the increments of the associated canonical statistic. Chapter 5 treats random processes whose likelihood functions belong to $(n, n-1)$-curved exponential families. Under very mild conditions, these are random-time transformations of Lévy processes. Chapter 10 can be viewed as an extension of Chapter 5; it treats the same processes and a larger class of stopping times that lead to exact or approximate Lévy processes. This makes classical results from sequential analysis applicable for such processes. Chapter 6 studies the exponential structure associated with underlying Markov processes.

The curved-exponential families associated with finite-time observations of a typical object from the exponential families of stochastic processes can always be extended to larger exponential families by suitable expansion of the parameter spaces. However, an extension of the associated random process to match such extended families is usually impossible. Chapter 7 discusses the problems that arise from such extensions and stochastic process interpretations of the latter. Chapter 8 discusses the general likelihood theory of exponential families of stochastic processes, and Chapter 9 is devoted to a particular random process. Chapter 11 covers the more advanced mathematical treatment of exponential families of stochastic processes, based on semimartingale theory. The final chapter, 12, reviews different definitions of exponential families of stochastic processes appearing in the statistical literature. One appendix contains the necessary tools from stochastic calculus required for a smooth reading of the later part of the text, and another appendix is devoted to the fundamental identity in sequential analysis.

This book is suitable for advanced graduate courses or for self-study by doctoral students. In both cases the required background is a graduate course in the theory of stochastic processes, including the basics of martingale theory, and knowledge of the basics from the classical theory of exponential families.